DreamMotion: Space-Time Self-Similar Score Distillation for Zero-Shot Video Editing

Hyeonho Jeong[1], Jinho Chang[1], Geon Yeong Park[2], and Jong Chul Ye[1,2]

[1] Kim Jaechul Graduate School of AI, KAIST, South Korea
[2] Dept. of Bio and Brain Engineering, KAIST, South Korea
{hyeonho.jeong, jinhojsk515, pky3436, jong.ye}@kaist.ac.kr
Project page: https://hyeonho99.github.io/dreammotion

Fig. 1: Zero-shot video editing results. The second row presents videos produced with our method with a non-cascaded video diffusion model, while those in the bottom row are from a cascaded model. For a full display of results, visit our *project page*.

Abstract. Text-driven diffusion-based video editing presents a unique challenge not encountered in image editing literature: establishing real-world motion. Unlike existing video editing approaches, here we focus on score distillation sampling to circumvent the standard reverse diffusion process and initiate optimization from videos that already exhibit natural motion. Our analysis reveals that while video score distillation can effectively introduce new content indicated by target text, it can also cause significant structure and motion deviation. To counteract this, we propose to match the space-time self-similarities of the original video and the edited video during the score distillation. Thanks to the use of score distillation, our approach is model-agnostic, which can be applied for both cascaded and non-cascaded video diffusion frameworks. Through extensive comparisons with leading methods, our approach demonstrates its superiority in altering appearances while accurately preserving the original structure and motion.

Keywords: Video Editing · Diffusion Models · Score Distillation

1 Introduction

Building upon the progress in diffusion models [12, 41, 44], the advent of large-scale text-image pairs [38] brought an unprecedented breakthrough in text-driven image generative tasks. In particular, real-world image editing has undergone significant evolution, supported by foundational Text-to-Image (T2I) diffusion models [30,35–37]. However, extending the success of diffusion-based image editing to video editing introduces a significant challenge: modeling temporally consistent, real-world motion throughout the reverse diffusion process.

Existing methods leveraging T2I diffusion models typically start by inflating attention layers to attend to multiple frames simultaneously [5, 6, 18, 19, 33, 50, 54, 55, 59]. Yet, this technique falls short of achieving smooth and complete motion, as it depends on the implicit preservation of motion through the inflated attention layers. As a result, a commonly adopted solution is to employ additional visual hints that explicitly guide the reverse diffusion process. One strategy is to use attention map guidance, for example, by injecting self-attention maps [4, 33] or manipulating cross-attentions [26]. Other works attempt to integrate the denoising process with spatially-aligned structural cues, like depth or edge maps. For example, pre-trained adapter networks such as ControlNet [58] or GLIGEN [24] have been transferred from image to video domain, achieving structure-consistent outputs [5, 15, 18, 59].

Even with the presence of pre-trained Text-to-Video (T2V) diffusion models, zero-shot video editing still poses a significant hurdle since publicly available T2V models [45, 49] lack sufficiently rich temporal priors to accurately depict real-world motion in the generated videos, as illustrated in Fig. 2. Thus, recent endeavors often adopt a self-supervised strategy of finetuning pre-trained model weights on the motion presented in an input video [17, 29, 53, 60, 61]. Whether employing T2I or T2V models, the conventional reverse diffusion process —beginning with standard Gaussian noise or, at most, inverted latent representations— struggles to reprogram complex, real-world motion, unless supplemented by additional visual conditions or by overfitting the spatial-temporal priors to a particular video.

To this end, we propose to diverge from the previous video editing literature. Our approach, DreamMotion, deliberately avoids the standard denoising process (ancestral sampling), and instead leverages the Score Distillation Sampling (SDS, [32]) grounded optimization to edit a video. Specifically, starting from an input video with temporally consistent, natural motion, we attempt to progressively modify the video's appearance while maintaining the integrity of the motion. In specific, our framework gradually injects target appearance to the video using Delta Denoising Score (DDS, [9]) gradients within T2V diffusion models. During this procedure, we filter the gradients with additional binary mask conditions to avoid blurriness and over-saturation. While this optimization effectively infuses the targeted appearance, it tends to accumulate struc-

Fig. 2: Ancestral sampling-based zero-shot video editing fails to capture complex, real-world motion in the generated videos.

tural errors, resulting in deviations in motion across the final output frames (see Fig. 3). To address this, we present self-similarity-based space-time regularization methods. More specifically, by aligning the spatial self-similarity of diffusion features between the original and edited videos, we preserve structure and motion integrity while seamlessly modifying the appearance. Furthermore, ensuring temporal self-similarity between the two features facilitates effective temporal smoothing, preventing potential distortions in areas subjected to optimization. Our methodology is applied to both cascaded and non-cascaded video diffusion models, showcasing its wide applicability across different video editing frameworks.

In summary, DreamMotion offers the following key contributions:

- A pioneering zero-shot framework that distills video score from text-to-video diffusion priors to inject target appearance.
- A novel space-time regularization that aligns spatial self-similarity to minimize structural deviations and temporal self-similarity to prevent distortions.
- Comprehensive validation of our approach across two distinct setups: non-cascaded and cascaded video diffusion frameworks.

2 Background

Diffusion Models Diffusion models [12, 41, 44] define the generative process as the reverse of the forward noising process. For clean data represented by $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x})$, the forward process gradually introduces Gaussian noise through Markov transition with conditional densities

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) &= \mathcal{N}(\boldsymbol{x}_t \mid \beta_t \boldsymbol{x}_{t-1}, (1-\beta_t)\boldsymbol{I}), \\ p(\boldsymbol{x}_t \mid \boldsymbol{x}_0) &= \mathcal{N}(\boldsymbol{x}_t \mid \sqrt{\bar{\alpha}}\boldsymbol{x}_0, (1-\bar{\alpha})\boldsymbol{I}), \end{aligned} \tag{1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^d$ is a noised latent representation at timestep t and the noise schedule β_t is a monotonically increasing sequence of t with $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$. Then, the objective of diffusion model training is to obtain a multi-scale

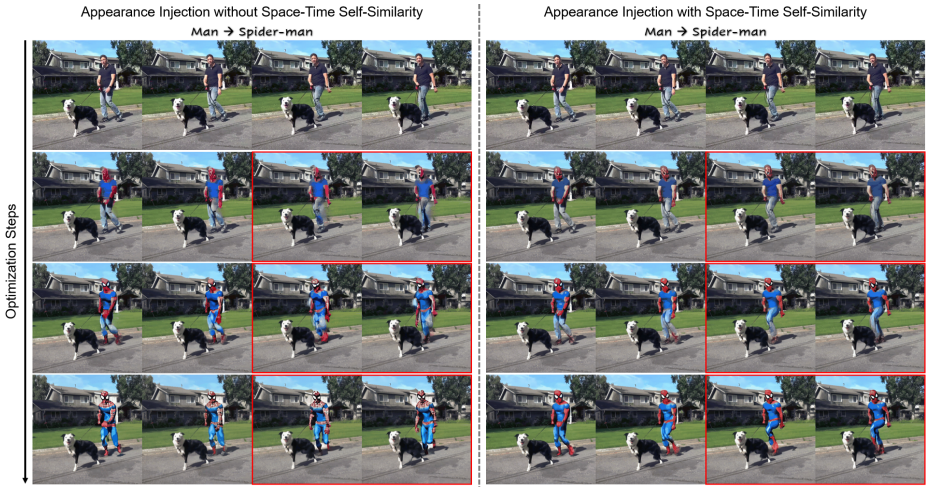

Fig. 3: Optimization progress visualization. The proposed self-similarity regularization effectively preserves the structure and motion of the original video.

U-Net denoiser $\boldsymbol{\epsilon}_{\phi^*}$ that satisfies

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{\boldsymbol{x}_t \sim p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0), \boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})} \left[\left\| \boldsymbol{\epsilon}_{\phi}(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon} \right\|_2^2 \right], \tag{2}$$

where $\boldsymbol{\epsilon}_{\phi^*}(\boldsymbol{x}_t, t) \simeq \boldsymbol{\epsilon} = \frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{\sqrt{1-\bar{\alpha}}}$. Notably, the Epsilon-Matching loss in (2) is equivalent to the Denoising Score Matching (DSM, [16, 43, 48]) with alternative parameterization:

$$\min_{\phi} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{s}_{\phi}^t(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) \right\|_2^2 \right], \tag{3}$$

where $\boldsymbol{s}_{\phi^*}(\boldsymbol{x}_t, t) \simeq -\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1-\bar{\alpha}} = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\phi^*}(\boldsymbol{x}_t, t)$. For the reverse process, with the learned noise prediction network $\boldsymbol{\epsilon}_{\phi}^*$, the noisy sample of previous timestep $\boldsymbol{x}_{t-1}$ can be estimated by:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\phi^*}(\boldsymbol{x}_t, t) \right) + \tilde{\beta}_t \boldsymbol{\epsilon}, \tag{4}$$

where $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$.

Conditional Generation In the context of conditional generation, data $\boldsymbol{x}$ is paired with an additional conditioning signal y, which in our case is a text caption. To train a text-driven diffusion model, the text conditional embedding y is incorporated into the objective as:

$$\min_{\phi} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}, y} \left[\left\| \boldsymbol{\epsilon}_{\phi}(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon} \right\| \right] \tag{5}$$

To augment the effect of text condition, classifier-free guidance [13] attempts to benefit from both conditional and unconditional noise prediction, using a single network. In specific, the epsilon prediction is defined as

$$\boldsymbol{\epsilon}_{\phi}^w(\boldsymbol{x}_t, t, y) = (1+w)\boldsymbol{\epsilon}_{\phi}(\boldsymbol{x}_t, t, y) - w\boldsymbol{\epsilon}_{\phi}(\boldsymbol{x}_t, t, \varnothing), \tag{6}$$

where $\varnothing$ denotes null text embedding and w is the guidance scale.

Video Diffusion Models Our framework leverages foundational video diffusion models for obtaining video scores. Consider a video sequence of N frames represented by $\boldsymbol{x}^{1:N} \in \mathbb{R}^{N\times d}$. For any n-th frame within this sequence, denoted by $\boldsymbol{x}^n \in \mathbb{R}^d$, the noisy frame latent $\boldsymbol{x}_t^n$ sampled from $p_t(\boldsymbol{x}_t^n|\boldsymbol{x}^n)$ can be expressed as $\boldsymbol{x}_t^n = \sqrt{\bar{\alpha}_t}\boldsymbol{x}^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^n$, where $\boldsymbol{\epsilon}_t^n \sim \mathcal{N}(0, I)$. Then, we similarly define $\boldsymbol{x}_t^{1:N}$ and $\boldsymbol{\epsilon}_t^{1:N}$. The objective of video diffusion model training is then to obtain a denoiser network ϵ_{ϕ^*} that satisfies:

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{\boldsymbol{x}_t^{1:N},\boldsymbol{x}^{1:N},\boldsymbol{\epsilon}^{1:N},y}\big[\, \big\|\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t^{1:N}, t, y) - \boldsymbol{\epsilon}^{1:N}\big\|\,\big], \tag{7}$$

where y is a text caption uniformly describing the video sequence $\boldsymbol{x}^{1:N}$.

Seeking to create videos that are both spatially and temporally enlarged and of high quality, video diffusion models have been expanded to cascaded pipelines [2, 11, 14, 40, 51, 57]. These cascaded video pipelines commonly follow a coarse-to-fine video generation approach, beginning with a module dedicated to creating keyframes that are low in both spatial and temporal resolution. Subsequent stages involve temporal interpolation and spatial super-resolution modules, which work to increase the temporal and spatial resolution of the frames, respectively. In this work, we plug our method into both cascaded and non-cascaded scenarios, proving its model-agnostic capability.

3 DreamMotion

3.1 Overview

Starting with a series of input video frames $\hat{\boldsymbol{x}}^{1:N}$, a corresponding text prompt $\hat{y}$, and a target text y, our goal is to get an edited video $\boldsymbol{x}^{1:N}$ that preserves the structural integrity and overall motion of $\hat{\boldsymbol{x}}^{1:N}$, while faithfully reflecting y. DreamMotion starts by initializing the target video variable $\boldsymbol{x}_0^{1:N}(\theta)$ by the original video $\hat{\boldsymbol{x}}^{1:N}$. Our optimization strategy is then three-pronged: (1) $\mathcal{L}_{\text{V-DDS}}$ that paints $\boldsymbol{x}_0^{1:N}(\theta)$ to match the appearance dictated by y, (2) $\mathcal{L}_{\text{S-SSM}}$ which encourages the structure of $\boldsymbol{x}_0^{1:N}(\theta)$ to align with $\hat{\boldsymbol{x}}^{1:N}$, (3) $\mathcal{L}_{\text{T-SSM}}$ which smoothens the gradients over the temporal dimension to eliminate any potential artifacts.

In Sec. 3.2, we briefly review SDS and DDS loss formulations and describe how we directly modify the appearance of $\boldsymbol{x}^{1:N}$ with DDS-based gradients. This technique, while effective in appearance injection, tends to accumulate structural inaccuracies, resulting in motion deviation in the end output. To address this, Sec. 3.3 introduces a strategy for structural correction based on self-similarity, and Sec. 3.4 details our approach for temporal smoothing, also leveraging self-similarity. Finally, in Sec. 3.5, we elaborate on the extension of DreamMotion to the cascaded video diffusion framework. For simplicity, we primarily describe the diffusion model as operating in pixel space throughout this paper. However, in practice, our implementation encompasses both a latent space-based (Sec. 4.1, [45]) and a pixel space-based video diffusion model (Sec. 4.2, [57]).

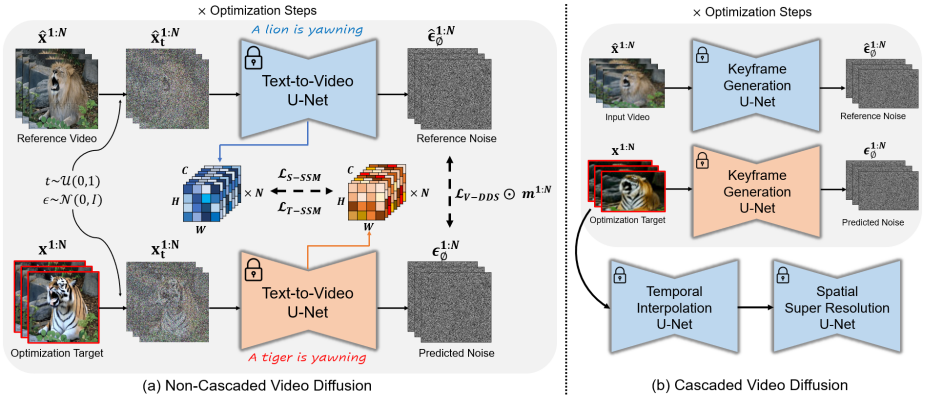

Fig. 4: Overview. DreamMotion leverages gradients derived from score distillation to inject target appearance, which is complemented by self-similarity alignments across spatial and temporal dimensions. This strategy seamlessly fits into cascaded video diffusion frameworks, confining the optimization on the keyframe generation phase.

3.2 Appearance Injection

Image Score Distillation Let $\boldsymbol{x}_0(\theta)$ denote the target image parameterized by θ and $\boldsymbol{\epsilon}_\phi$ represent a T2I diffusion model. SDS aims to align $\boldsymbol{x}_0(\theta)$ with the target text y by optimizing the diffusion training loss gradient, expressed as:

$$\mathcal{L}_{\text{SDS}}(\theta; y) = \left\| \boldsymbol{\epsilon}_\phi^w(\boldsymbol{x}_t(\theta), t, y) - \boldsymbol{\epsilon} \right\|_2^2, \tag{8}$$

with $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ and $t \sim \mathcal{U}(0,1)$. Although $\nabla_\theta \mathcal{L}_{\text{SDS}}$ provides an efficient gradient term for incrementally refining the image fidelity to the text y, SDS often results in over-saturation, blurriness, and lack of details in the generated image [9, 20, 25, 28, 52].

Under the assumption that the SDS score should be zero for pairs of correctly matched prompts and images, DDS [9] enhances the gradient direction obtained from the SDS framework by incorporating an additional text-image pair, comprising a reference text $\hat{\boldsymbol{y}}$ and a reference image $\hat{\boldsymbol{x}}_0$. Specifically, the noisy direction of the SDS score is calculated using the reference text-image branch, and this noisy score is subtracted from the main SDS optimization branch:

$$\mathcal{L}_{\text{DDS}}(\theta; y) = \left\| \boldsymbol{\epsilon}_\phi^w(\boldsymbol{x}_t(\theta), t, y) - \boldsymbol{\epsilon}_\phi^w(\hat{\boldsymbol{x}}_t, t, \hat{y}) \right\|_2^2. \tag{9}$$

Video Score Distillation with Masked Gradients Leveraging a pre-trained T2V diffusion model $\boldsymbol{\epsilon}_\phi$, we extend the DDS mechanism to distill video scores. Let $\boldsymbol{x}_0^{1:N}(\theta)$ represent the target video parameterized by θ, and $\boldsymbol{x}_0^{1:N}$ denote the fixed, source video. We optimize the video variable $\boldsymbol{x}_0^{1:N}(\theta)$ to reflect target text y by minimizing:

$$\mathcal{L}_{\text{V-DDS}}(\theta; y) = \left\| \boldsymbol{\epsilon}_\phi^w(\boldsymbol{x}_t^{1:N}(\theta), t, y) - \boldsymbol{\epsilon}_\phi^w(\hat{\boldsymbol{x}}_t^{1:N}, t, \hat{y}) \right\|_2^2. \tag{10}$$

While the video delta denoising score (V-DDS) offers a reliable gradient for gradually injecting appearance described by target text y, it still suffers from

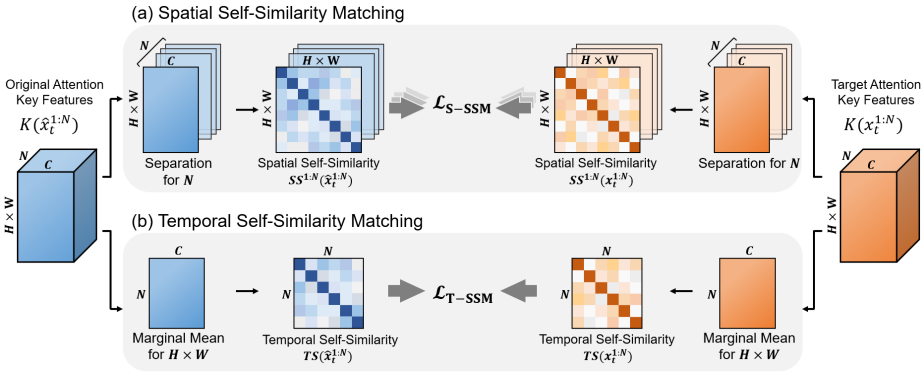

Fig. 5: The proposed space-time self-similarity regularization: (a) Spatial Self-Similarity Matching and (b) Temporal Self-Similarity Matching

blurriness and over-saturation. We mitigate this issue by additional mask conditioning. Specifically, we filter the obtained gradients with a sequence of masks $m^{1:N}$ that annotate the objects to be edited in each frame, by $\nabla_\theta \mathcal{L}_{\text{V-DDS}} \odot m^{1:N}$. The filtered gradients ensure that unintended regions in $\boldsymbol{x}_0^{1:N}(\theta)$ remain unaffected during V-DDS optimization (see Fig. 6).

A more significant issue arises when inaccurate gradients of $\mathcal{L}_{\text{V-DDS}}$ accumulate structural errors throughout the optimization process. Unlike editing still images, these errors are particularly problematic in video editing, as their accumulation deters temporal consistency within frames and often results in motion deflection, as illustrated in Fig. 3, 9. To tackle this, we propose to match self-similarities between target and reference branches, as detailed in Section 3.3.

3.3 Structure Correction

Spatial Self-Similarity Matching To address structural integrity, we require a representation that remains resilient against local texture patterns while retaining the global layout and overall shape of objects: self-similarity descriptors. Self-similarity of visual features facilitates identifying objects by emphasizing the relationship of an object's appearance to its surroundings, rather than relying on its absolute appearance. This principle of relative appearance has been effectively applied across various domains: in traditional methods for matching visual patterns [39], in the realm of neural style transfer through deep convolutional neural network features [21], and more recently, in the field of image editing utilizing DINO ViT features [3, 22, 47].

Our contribution lies in pioneering the application of self-similarity through deep diffusion features [46] to ensure structural correspondence between the target video $\boldsymbol{x}^{1:N}$ and the original video $\hat{\boldsymbol{x}}^{1:N}$. To achieve this, we add *identical noise* of timestep t to both videos (Eq. 1), resulting in $\boldsymbol{x}_t^{1:N}$ and $\hat{\boldsymbol{x}}_t^{1:N}$, which are then feed-forwarded to the video diffusion U-Net $\boldsymbol{\epsilon}_\phi$ to extract a pair of attention key features $K(\boldsymbol{x}_t^{1:N}), K(\hat{\boldsymbol{x}}_t^{1:N}) \in \mathbb{R}^{N \times (H \times W) \times C}$. Subsequently, we calculate spatial

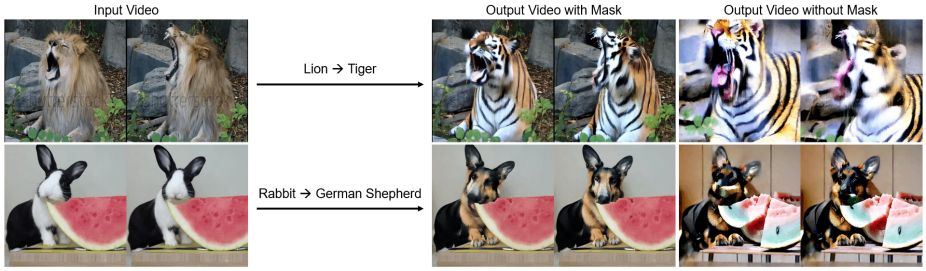

Fig. 6: Filtering optimization gradients plays a crucial role in maintaining visual fidelity and preserving the structure of the input video. Bounding boxes detected by off-the-shelf models [23, 27] are used to create binary masks indicating the target regions for editing.

self-similarity map $SS^n(\cdot) \in \mathbb{R}^{(H\times W)\times(H\times W)}$ of each n-th frame as follows:

$$SS^n_{i,j}(\boldsymbol{x}_t^{1:N}) = cos(K_i^n(x_t^{1:N}), K_j^n(x_t^{1:N})), \tag{11}$$

where $cos(\cdot,\cdot)$ denotes the normalized cosine similarity, i, j are all pairs of spatial indexes ($1 \le i, j \le (H\times W)$), and $\boldsymbol{x}_t^{1:N}(\theta)$ is simplified to $\boldsymbol{x}_t^{1:N}$ for brevity. The spatial self-similarity matching objective is formulated as:

$$\mathcal{L}_{\text{S-SSM}}(\boldsymbol{x}_t^{1:N}, \hat{\boldsymbol{x}}_t^{1:N}) = \frac{1}{N}\sum_{n=1}^{N}\left\| SS^n(\boldsymbol{x}_t^{1:N}) - SS^n(\hat{\boldsymbol{x}}_t^{1:N})\right\|_2^2, \tag{12}$$

thereby quantifying and minimizing the discrepancy between the self-similarity maps of the target and original videos.

3.4 Temporal Smoothing

Temporal Self-Similarity Matching Although the spatial self-similarity alignment, facilitated by $\mathcal{L}_{\text{S-SSM}}$, proficiently maintains structural consistency between the original and modified videos, it operates as a frame-independent optimization method, without considering the temporal correlation between frames. As observed in Fig. 9, such per-frame operations can lead to localized distortions and notable flickering in the optimized frames. To address these artifacts, we introduce a temporal regularization of $\mathcal{L}_{\text{S-SSM}}$ that models temporal correlations by leveraging self-similarity along the frame axis.

Calculating self-similarity over time necessitates a method to compress spatial information while retaining essential spatial details. For this purpose, we employ spatial marginal mean a first-order statistic, spatial marginal mean, as our global descriptor. This choice is supported by prior works [21,56], which have demonstrated their effectiveness in capturing crucial spatial details and serving as a robust global descriptor. More concretely, we condense the spatial dimensions of the extracted key features $K(\boldsymbol{x}_t^{1:N}) \in \mathbb{R}^{N\times(H\times W)\times C}$ to $M[K(\boldsymbol{x}_t^{1:N})] \in \mathbb{R}^{N\times C}$ through the process defined as:

$$M[K(\boldsymbol{x}_t^{1:N})] = \frac{1}{H\cdot W}\sum_{i=1}^{H\cdot W} K_i(\boldsymbol{x}_t^{1:N}), \tag{13}$$

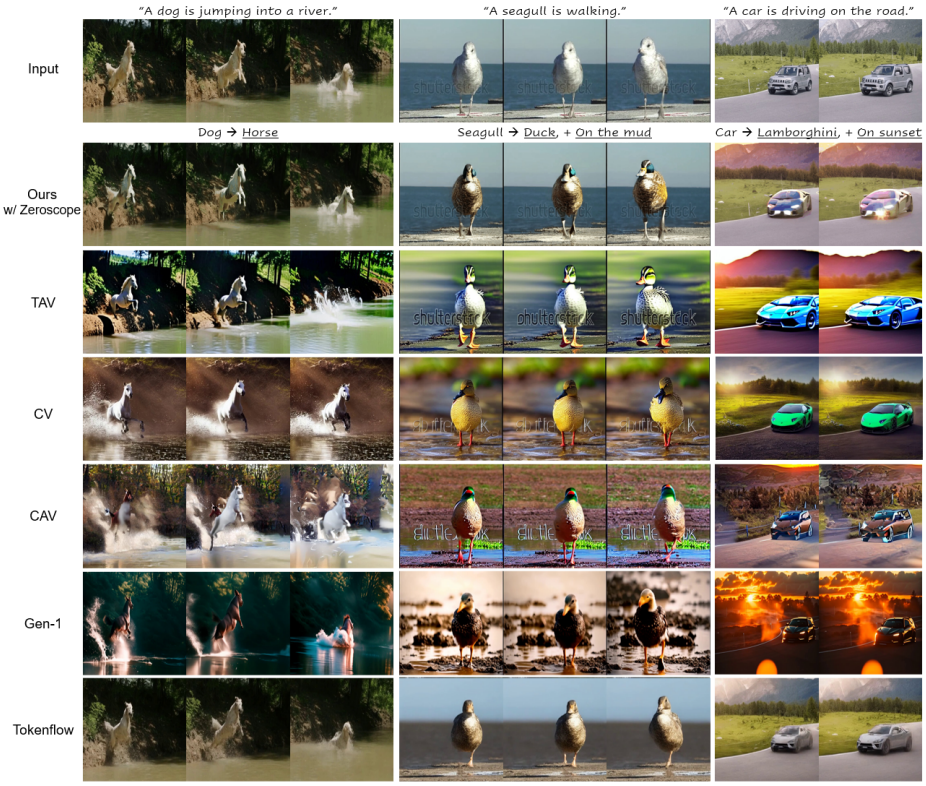

Fig. 7: Comparison. DreamMotion, applied to the Zeroscope model, is evaluated against five baseline methods. For a detailed assessment, please visit our project page.

where H and W denote the height and width, respectively, and C represents the channel dimension of the feature maps. We then establish the temporal self-similarity $TS(\cdot) \in \mathbb{R}^{N \times N}$ as follows:

$$TS_{i,j}(\boldsymbol{x}_t^{1:N}) = cos(M_i[K(\boldsymbol{x}_t^{1:N})], M_j[K(\boldsymbol{x}_t^{1:N})]), \tag{14}$$

where i, j are from frame indexes ($1 \leq i, j \leq N$). Subsequently, the temporal self-similarity matching loss is formulated as:

$$\mathcal{L}_{\text{T-SSM}}(\boldsymbol{x}_t^{1:N}, \hat{\boldsymbol{x}}_t^{1:N}) = \left\| TS(\boldsymbol{x}_t^{1:N}) - TS(\hat{\boldsymbol{x}}_t^{1:N}) \right\|_2^2. \tag{15}$$

It's noteworthy that the three losses $\mathcal{L}_{\text{V-DDS}}$, $\mathcal{L}_{\text{S-SSM}}$ and $\mathcal{L}_{\text{T-SSM}}$ share the same noise $\boldsymbol{\epsilon}$ and time t for their computations, achieving a computationally efficient integration of optimizations through a single forward and reverse diffusion step.

3.5 Expansion to Cascade Video Diffusion

As outlined in Section 2, cascaded video diffusion models commonly utilize a coarse-to-fine approach for video generation, comprising three specialized modules that function in sequence: Keyframe Generation, Temporal Interpolation,

and Spatial Super Resolution. Rather than applying the optimization process through this comprehensive pipeline—a process that would result in prohibitively high computational costs—we focus our efforts exclusively on the initial Keyframe Generation stage. Within this approach, we reinterpret $\boldsymbol{x}^{1:N}$ and $\hat{\boldsymbol{x}}^{1:N}$ to represent, respectively, the target and original keyframes, both resized to accommodate the low-resolution requirements of the keyframe generation space. Furthermore, we designate $\boldsymbol{\epsilon}_\phi$ to represent the keyframe generation U-Net, excluding the temporal interpolation and super-resolution modules. Following this setup, we apply our optimizations—$\mathcal{L}$V-DDS, $\mathcal{L}$S-SSM, and $\mathcal{L}$T-SSM—directly to $\boldsymbol{x}^{1:N}$. After completing the optimization, these refined keyframes undergo further processing through the Temporal Interpolation and Spatial Super Resolution stages. This comprehensive procedure is depicted in Fig. 4-(b), illustrating the streamlined approach to integrating our optimization methods within the cascaded video diffusion model framework.

4 Experiments

4.1 Non-cascaded Video Diffusion Framework

Setup For evaluation, we chose 26 text-video pairs from the public DAVIS [31] and WebVid [1] datasets. The videos vary in length from 8 frames to 16 frames. In this experiment, we deploy our method on ZeroScope [45], a foundational text-to-video latent diffusion model. The CFG scale w is configured as 9.0. We perform optimization for 200 steps using stochastic gradient descent (SGD) with a learning rate of 0.4. The optimization of an 8-frame video requires approximately 2 minutes, while optimizing a 16-frame video takes around 4 minutes, utilizing a single A100 GPU.

Baselines Our method is evaluated alongside 1 one-shot and 4 zero-shot video editing baselines. Tune-A-Video (TAV, [54]) selectively finetunes attention projection layers within an inflated T2I model on the given input video. ControlVideo (CV, [59]) integrates temporally extended ControlNet [58] to T2I diffusion and achieves motion-consistent video generation without any finetuning. Both Control-A-Video (CAV, [5]) and Gen-1 [7] are video diffusion models trained on large-scale text-image and text-video data. They explicitly guide the ancestral denoising process with a series of structural conditions like depth maps. Tokenflow [8] accomplishes time-consistent video editing by enforcing uniformity on the internal diffusion features across frames, in a zero-shot manner.

Qualitative Results Fig. 7 offers a qualitative comparison between our method and state-of-the-art baselines; for complete videos, refer to our *project page*. Our method produces temporally consistent videos that closely adhere to the target prompt while most accurately preserving the motion of the input video, a feat that other baselines struggle to achieve simultaneously.

Quantitative Results We conducted a comprehensive quantitative evaluation, which includes both automatic metrics and a user study. The summarized results can be found in Tab. 1.

(a) Automatic metrics. We first employ CLIP [34] to measure the text alignment and frame consistency of the edited videos. For assessing textual alignment [10], we measure average cosine similarity between the target text prompt and the edited frames. In terms of frame consistency, we calculate CLIP image features for every frame in the output video and then compute the average cosine similarity across all neighboring pairs of frames. We additionally compute tracking-based motion fidelity score [56] and framewise LPIPS [26] for measuring spatial consistency. According to the results in Tab. 1, our approach surpasses the baselines in achieving higher textual alignment and better spatial-temporal consistency.

(b) User study. We surveyed 36 participants to assess the accuracy of editing, temporal consistency, and preservation of structure & motion, using a rating scale from 1 to 5. Participants were shown the input video followed by anonymized output videos from each baseline. They were then asked the three questions: (i) Edit Accuracy: Does the output video accurately reflect the target text by appropriately editing all relevant elements? (ii) Frame Consistency: Are the frames in the output video temporally consistent? (iii) Structure and Motion Preservation: Has the structure and motion of the input video been accurately maintained in the output video? Tab. 1 illustrates that our method outperforms the baselines in all measured aspects.

	Automatic Metrics				Human Evaluation		
Method	Text-Align	Frame-Con	Motion-Fidelity	Frame-LPIPS	Edit-Acc	Frame-Con	SM-Preserve
Tune-A-Video	0.8177	0.9218	0.6947	0.4172	3.52	2.82	2.89
ControlVideo	0.7850	0.9678	-	0.3763	2.74	2.68	2.03
Control-A-Video	0.7848	0.9297	0.8453	0.3829	2.17	2.16	2.18
Gen-1	<u>0.8192</u>	<u>0.9704</u>	-	-	3.31	<u>3.62</u>	2.95
Tokenflow	0.7813	0.9576	<u>0.9184</u>	<u>0.3427</u>	<u>3.63</u>	3.54	<u>3.92</u>
Ours (Zeroscope)	**0.8209**	**0.9726**	**0.9259**	**0.3042**	**4.14**	**4.21**	**4.33**

Table 1: Quantitative evaluations. DreamMotion with Zeroscope outperforms various video editing methods in all seven features.

4.2 Cascaded Video Diffusion Framework

Setup In this experiment, we utilize the 8-frame videos from the previously assembled text-video pairs. Additionally, we benefit from Show-1 [57], an open-source, cascaded video diffusion model. As detailed in Sec. 3.5, we compose our cascaded pipeline comprising Keyframe Generation, Temporal Interpolation, and Spatial Super Resolution, with all modules operating in pixel space. Our method is implemented during the initial keyframe generation stage. During keyframe optimization, these input videos undergo resizing to a resolution of 80x128 pixels,

Fig. 8: Comparison. DreamMotion with Show-1 cascaded model is evaluated against two baselines.

	Automatic Metrics		Human Evaluation		
Method	Text-Align	Frame-Con	Edit-Acc	Frame-Con	SM-Preserve
Inversion + Word Swap	0.7586	0.9714	3.36	3.42	2.21
VMC	0.7563	0.9703	3.13	3.22	3.35
Ours (Show-1)	**0.7747**	**0.9755**	**3.97**	**3.74**	**4.30**

Table 2: Quantitative evaluations. DreamMotion utilizing Show-1 surpasses other cascaded baselines across the five features. Other baselines were also implemented using the same video model, ensuring a fair comparison.

with the optimization process taking approximately 3 minutes on a single A100 GPU. Following the optimization, the frame interpolation and super-resolution modules expand the output keyframes temporally and spatially, respectively.

Baselines To our knowledge, VMC [17] stands out as the sole video editing approach utilizing a cascaded video diffusion pipeline. VMC adapts temporal attention layers within the keyframe generation module, leveraging their novel motion distillation objective. For comparison purposes, we introduce an additional variant that employs direct inference using the cascaded pipeline with modified target text, starting from the DDIM inverted latents [42].

Qualitative Results We qualitatively compare our method against baselines in Fig. 8. DreamMotion generates videos that match the structure and layout of the input video while adhering to the edit prompt, while other methods struggle to maintain the structural and motion integrity of the original video. Since all three methods use unaltered temporal interpolation and super-resolution models after the generation of keyframes, they commonly produce temporally consistent videos. For comprehensive results, please refer to the appendix.

Fig. 9: Ablation of spatial and temporal self-similarity alignments. Joint optimization of $\mathcal{L}_{\text{V-DDS}} + \mathcal{L}_{\text{S-SSM}} + \mathcal{L}_{\text{T-SSM}}$ generates the optimal output videos.

Quantitative Results Adopting the metrics outlined in Sec. 4.1, we compare our method quantitatively against baseline approaches, detailed in Tab. 2. Notably, our approach demonstrated substantial superiority in Structure and Motion Preservation (SM-Preserve).

4.3 Ablation Studies

In Fig. 6, we evaluate the impact of using bounding box-driven masks to selectively filter gradients during $\mathcal{L}_{\text{V-DDS}}$ update. The results demonstrate that filtering gradients responsible for appearance injection enhances the precision of video editing and improves visual fidelity while avoiding issues of blurriness and saturation.

We next ablate the necessity of our self-similarity guidances. Fig. 3 illustrates the optimization progress with and without our self-similarity alignments. The process begins with the initial input video (top row). Solely using $\mathcal{L}_{\text{V-DDS}}$ for appearance injection (left) leads to the accumulation of structural errors as optimization progresses, resulting in motion deviation in the final output. However, when the process is regularized by the spatial and temporal self-

	Text-Align	Frame-Con	Motion-Fidelity	Frame-LPIPS
Ours *wo* $\mathcal{L}_{S\text{-}SSM} + \mathcal{L}_{T\text{-}SSM}$	0.8202	0.9648	0.8426	0.3247
Ours *wo* $\mathcal{L}_{T\text{-}SSM}$	0.8114	0.9567	0.9011	0.3186
Ours *wo masks*	0.8180	0.9695	0.8653	0.3416
Ours *(full)*	**0.8209**	**0.9726**	**0.9259**	**0.3042**

Table 3: Quantitative ablation. We demonstrate the impact of each factor by removing individual losses and masking conditions.

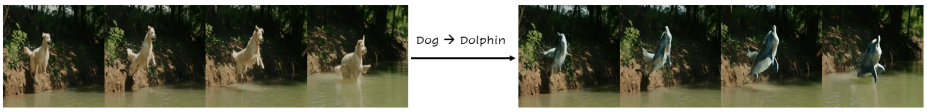

Fig. 10: Limitation. DreamMotion limits its ability to produce videos that necessitate substantial structural alterations.

similarities (right), edited videos maintain the structure and motion fidelity throughout the optimization. Additionally, in Fig. 9, we illustrate video editing results under different optimization setups: (i) $\mathcal{L}_{\text{V-DDS}}$. (ii) $\mathcal{L}_{\text{V-DDS}} + \mathcal{L}_{\text{S-SSM}}$. (ii) $\mathcal{L}_{\text{V-DDS}} + \mathcal{L}_{\text{S-SSM}} + \mathcal{L}_{\text{T-SSM}}$. The absence of spatial self-similarity loss leads to inconsistency in object structures across frames. For instance, the shape of a bird's wing varies, creating visible discrepancies, as shown in Fig. 9-*left*. While aligning spatial self-similarity with the original video preserves structural integrity, it may generate artifacts in optimized areas. However, these artifacts are efficiently addressed through the addition of temporal self-similarity guidance. Lastly, Tab. 3 provides a quantitative analysis of each optimization term and masking condition.

5 Conclusion

In this work, we have addressed the intricate challenge of diffusion-based video editing, a domain where formulating temporally consistent, real-world motion remains a notable obstacle. DreamMotion introduced score distillation-based optimization to text-to-video diffusion models, marking a departure from traditional, ancestral sampling-based video editing. Our framework adeptly incorporated new content as specified by target text descriptions using the Video Delta Denoising Score, while preserving the the structural integrity and motion of the original video via a novel space-time self-similarity alignment. Through rigorous validation in both cascaded and non-cascaded video diffusion settings, our approach has proven superior in maintaining the essence of the original video while seamlessly integrating desired alterations. Regarding limitations, our framework is designed to preserve the structural integrity of the original video, and as such, it is not suited for edits that require significant structural changes (see Fig. 10).
Ethics Statement Our work is based on generative models that carry the risk of being repurposed for unethical uses, such as misleading content.

Acknowledgments

This work was supported by the National Research Foundation of Korea under Grant RS-2024-00336454. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, Ministry of Science and ICT) (No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)). This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023. This research was supported by Field-oriented Technology Development Project for Customs Administration funded by the Korea government (the Ministry of Science & ICT and the Korea Customs Service) through the National Research Foundation (NRF) of Korea under Grant NRF2021M3I1A1097910.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
5. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
6. Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: Flatten: optical flow-guided attention for consistent text-to-video editing. arXiv preprint arXiv:2310.05922 (2023)
7. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
8. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
9. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023)

10. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
11. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)
15. Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. arXiv preprint arXiv:2307.14073 (2023)
16. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**(4) (2005)
17. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9212–9221 (2024)
18. Jeong, H., Ye, J.C.: Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In: The Twelfth International Conference on Learning Representations
19. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
20. Kim, S., Lee, K., Choi, J.S., Jeong, J., Sohn, K., Shin, J.: Collaborative score distillation for consistent visual synthesis. arXiv preprint arXiv:2307.04787 (2023)
21. Kolkin, N., Salavon, J., Shakhnarovich, G.: Style transfer by relaxed optimal transport and self-similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10051–10060 (2019)
22. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264 (2022)
23. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
24. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
25. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
26. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

28. Nam, H., Kwon, G., Park, G.Y., Ye, J.C.: Contrastive denoising score for text-guided latent diffusion image editing. arXiv preprint arXiv:2311.18608 (2023)
29. Park, G.Y., Jeong, H., Lee, S.W., Ye, J.C.: Spectral motion alignment for video motion transfer using diffusion models. arXiv preprint arXiv:2403.15249 (2024)
30. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
31. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
32. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
33. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
37. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
38. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
39. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
40. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
41. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
43. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
44. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
45. Sterling, S.: Zeroscope (2023), `https://huggingface.co/cerspense/zeroscope_v2_576w`

46. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36** (2024)
47. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10748–10757 (2022)
48. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation **23**(7), 1661–1674 (2011)
49. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
50. Wang, W., Jiang, Y., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
51. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
52. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36** (2024)
53. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. arXiv preprint arXiv:2312.04433 (2023)
54. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
55. Wu, R., Chen, L., Yang, T., Guo, C., Li, C., Zhang, X.: Lamp: Learn a motion pattern for few-shot-based video generation. arXiv preprint arXiv:2310.10769 (2023)
56. Yatim, D., Fridman, R., Tal, O.B., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. arXiv preprint arXiv:2311.17009 (2023)
57. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
58. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
59. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
60. Zhang, Y., Tang, F., Huang, N., Huang, H., Ma, C., Dong, W., Xu, C.: Motioncrafter: One-shot motion customization of diffusion models. arXiv preprint arXiv:2312.05288 (2023)
61. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)